

Pixel Objectness

Suyog Dutt Jain and Bo Xiong and Kristen Grauman

Department of Computer Science

The University of Texas at Austin

<http://vision.cs.utexas.edu/projects/pixelobjectness/>

We propose an end-to-end learning framework for foreground object segmentation. Given a single novel image, our approach produces a pixel-level mask for all “object-like” regions—even for object categories never seen during training. We formulate the task as a structured prediction problem of assigning a foreground/background label to each pixel, implemented using a deep fully convolutional network. Key to our idea is training with a mix of *image-level* object category examples together with relatively few images with *boundary-level* annotations. Our method substantially improves the state-of-the-art on foreground segmentation on the ImageNet and MIT Object Discovery datasets—with 19% absolute improvements in some cases. Furthermore, on over 1 million images, we show it generalizes well to segment object categories unseen in the foreground maps used for training. Finally, we demonstrate how our approach benefits image retrieval and image retargeting, both of which flourish when given our high-quality foreground maps.

I. INTRODUCTION

Foreground object segmentation is a fundamental vision problem with a wide variety of applications. For example, a visual search system can use foreground segmentation to focus on the important objects in the query image, ignoring background clutter that can adversely affect the search. Object segmentation is also a prerequisite in graphics applications like rotoscoping, image retargeting, and 3d-reconstruction. Knowing the spatial extent of objects can also benefit downstream vision tasks like scene understanding, caption generation, and summarization.

In any such setting, it is crucial to segment “generic” objects in a *category-independent* manner. That is, the system must be able to identify object boundaries for objects it has never encountered during training.¹

Today there are two main strategies for generic object segmentation: saliency and object proposals. Both strategies capitalize on properties that can be learned from images and generalize to unseen objects (e.g., well-defined boundaries, differences with surroundings, shape cues, etc.).

Saliency methods identify regions likely to capture human attention. They yield either highly localized attention maps [5], [6], [7] or a complete segmentation of the prominent object [8], [9], [10], [11], [12]. Saliency focuses on regions that stand out, which is not the case for all foreground objects.

Alternatively, *object proposal* methods learn to localize all objects in an image, regardless of their category [13], [14],

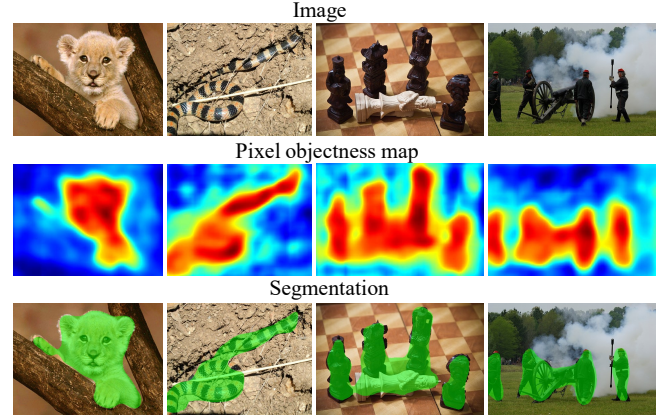


Fig. 1: Our goal is to predict an objectness map for each pixel (2nd row) and a single foreground segmentation (3rd row). Left to right: Our method can accurately handle objects with occlusion, thin objects with similar colors to background, man-made objects, and multiple objects. Our approach is class-independent, meaning it is *not* trained to detect the particular objects in the images (best viewed on pdf).

[15], [16], [17], [18]. Proposal methods aim to obtain high recall at the cost of low precision, i.e., they must generate a large number of object proposals (typically 1000s) to accurately cover all objects in an image. This usually involves a multi-stage process: first bottom-up segments are extracted, then they are scored by their degree of “objectness”. Relying on bottom-up segments can be limiting, since low-level cues may fail to pull out contiguous regions for more complex objects. Furthermore, in practice, the scores produced by proposal methods are not so reliable such that one can rely exclusively on the top few proposals.

Motivated by these shortcomings, we introduce *pixel objectness*, a new approach to generic foreground segmentation. Given a novel image, the goal is to determine the likelihood that each pixel is part of a foreground object (as opposed to background or “stuff” classes like grass, sky, sidewalks, etc.) Our definition of a generic foreground object follows that commonly used in the object proposal literature [13], [14], [15], [16], [17], [18]. Pixel objectness quantifies how likely a pixel belongs to an object of *any* class, and should be high even for objects unseen during training. See Fig. 1.

We cast foreground object segmentation as a unified structured learning problem, and implement it by training a deep fully convolutional network to produce dense (binary) pixel label maps. Given the goal to handle arbitrary objects, one might expect to need ample foreground-annotated examples across a vast array of categories to learn the generic cues. However, we show that, somewhat surprisingly, when training

¹This differentiates the problem from traditional recognition or “semantic segmentation” [1], [2], [3], [4], where the system is trained specifically to find predefined categories, and is not equipped to segment any others.

with *explicit boundary-level* annotations for few categories pooled together into a single generic “object-like” class, pixel objectness generalizes well to *thousands* of unseen objects. This generalization ability is facilitated by an *implicit image-level* notion of objectness built into a pretrained classification network, which we transfer to our segmentation model during initialization.

Our formulation has some key advantages. First, our method is not limited to segmenting objects that stand out conspicuously, as is typically the case in salient object detection [9], [10], [11], [12]. Second, it is not restricted to segmenting only a fixed number of object categories, as is the case for supervised semantic segmentation [1], [2], [3], [4]. Third, rather than divide processing into two independent steps—as is typical in today’s region-based object proposals [13], [14], [15], [16], [17], [18]—our method unifies learning “what makes a good region” with learning “which pixels belong in a region together”. As such, unlike the two-stage proposal methods, it is not beholden to whatever flawed regions some bottom-up segmenter might produce.

Through extensive experiments, we show that our model generalizes very well to unseen objects. We obtain state-of-the-art foreground object segmentations on the challenging ImageNet [19] and MIT Object Discovery [20] datasets. Our gains are often significant (e.g., 19% absolute increase in overlap scores) on these well-studied benchmarks, which have recently seen only incremental progress. Finally, we show how to leverage our segmentations to benefit object-centric image retrieval and content-aware image resizing. Please see our project webpage² for code and pre-trained models.

II. RELATED WORK

We divide related work into two top-level groups: (1) methods that extract an object mask no matter the object category, and (2) methods that learn from category-labeled data, and seek to recognize/segment those particular categories in new images. Our method fits in the first group.

A. Category-independent segmentation

Interactive image segmentation algorithms such as the popular GrabCut [21] let a human guide the algorithm using bounding boxes or scribbles. These methods are most suitable when high precision segmentations are required such that some guidance from humans is worthwhile. While some methods try to minimize human involvement [22], [23], still typically a human is always in the loop to guide the algorithm. In contrast, our model is fully automatic and segments foreground objects without any human guidance.

Object proposal methods, also discussed above, produce thousands of generic object proposals either in the form of bounding boxes [16], [17], [18], [24] or object regions [13], [14], [15]. Generating thousands of hypotheses ensures high recall, but at the same time often results in very low precision. Though effective for object detection, it is difficult to automatically filter out accurate proposals from this large hypothesis

set without class-specific knowledge. We instead generate a *single* hypothesis of the foreground as our final segmentation. Our experiments directly evaluate our method’s advantage.

Saliency models have also been widely studied [11], [12], [8] including methods based on deep learning [7], [6], [5]. The goal is to identify regions that are likely to capture human attention. While some methods produce highly localized regions (e.g. just the face of a person), others segment complete objects [9], [10], [11], [12]. While saliency focuses on objects that “stand out”, our method is designed to segment all foreground objects, irrespective of whether they stand out in terms of low-level saliency.

B. Category-specific segmentation

Semantic segmentation refers to the task of jointly *recognizing* and segmenting objects, classifying each pixel into one of k fixed categories. Recent advances in deep learning have fostered increased attention to this task. Most deep semantic segmentation models include fully convolutional networks that apply successive convolutions and pooling layers followed by upsampling or deconvolution operations in the end to produce pixel-wise segmentation maps [1], [2], [3], [4]. However, these methods are trained for a fixed number of categories. We are the first to show that a fully convolutional network can be trained to accurately segment *arbitrary* foreground objects. Though relatively few categories are seen in training, our model generalizes very well to unseen categories (as we demonstrate for 3,624 classes from ImageNet, only a fraction of which overlap with PASCAL, the source of our training masks).

Weakly supervised joint segmentation methods use weaker supervision than semantic segmentation methods. Given a batch of images known to contain the same object category, they segment the object in each one [25], [26], [27], [28], [20], [29]. The idea is to exploit the similarities in the batch of related images (e.g., using region or pixelwise dense matching) to discover the common foreground object. The output is either a pixel-level mask [25], [30], [27], [28], [31], [20], [29] or bounding box [26], [32]. While joint segmentation is useful, its performance is limited by the shared structure that exists in the collection; a diverse collection with different shapes and viewpoints poses a significant challenge to existing methods. Moreover, in most practical scenarios, such weak supervision is not available. A stand alone single-image segmentation model like ours is more widely applicable.

Propagation-based methods transfer information from exemplars with human-labeled foreground masks [33], [34], [35], [36], [37]. They usually involve a matching stage between likely foreground regions and the exemplars. Foreground regions from strong matches are then transferred to obtain the final segmentation. The downside is the need to store a large amount of exemplar data at test time and perform an expensive and potentially noisy matching process for each test image. In contrast, our segmentation model, once trained end-to-end, is very efficient to apply and does not need to retain any training data.

²<http://vision.cs.utexas.edu/projects/pixelobjectness/>

III. APPROACH

Our goal is to design a model that can predict the likelihood of each pixel being a generic foreground object as opposed to background. We refer to our task as *pixel objectness*. We use this name to distinguish our task from the related problems of salient object detection (which seeks only the most attention-grabbing foreground object) and region proposals (which seeks a ranked list of candidate object-like regions). We pose pixel objectness as a dense labeling problem, and propose a solution based on a convolutional neural network architecture that supports end-to-end training.

First we introduce our core approach and describe the model architecture (Sec. III-A). Then, we explore two applications that illustrate the utility of pixel objectness—content-based image retrieval and image retargeting—both of which demand a single, high-quality estimate of the foreground object region(s) (Sec. III-B).

A. Predicting Pixel Objectness

Problem formulation: Given an RGB image \mathcal{I} of size $m \times n \times c$ as input, we formulate the task of foreground object segmentation as densely labeling each pixel in the image as either “object” or “background”. Thus the output of pixel objectness is a binary map of size $m \times n$.

Since our goal is to predict objectness for each pixel, our model should 1) predict a pixel-level map that aligns well with object boundaries, and 2) generalize so it can assign high probability to pixels of unseen object categories.

Challenges in dense foreground-labeled training data: Potentially, one way to address both challenges would be to rely on a large annotated image dataset that contains a large number of diverse object categories with pixel-level foreground annotations. However, such a dataset is non-trivial to obtain. The practical issue is apparent looking at recent large-scale efforts to collect segmented images. They contain boundary-level annotations for merely dozens of categories (20 in PASCAL [38], 80 in COCO [39]), and/or for only a tiny fraction of all dataset images (0.03% of ImageNet’s 14M images have such masks). Furthermore, such annotations come at a price—about \$400,000 to gather human-drawn outlines on 2.5M object instances from 80 categories [39] assuming workers receive minimum wage. To naively train a *generic* foreground object segmentation system, one might expect to need foreground labels for many more representative categories, suggesting an alarming start-up annotation cost.

Mixing explicit and implicit representations of objectness: This challenge motivates us to consider a different means of supervision to learn generic pixel objectness. Our idea is to train the system to predict pixel objectness using a mix of *explicit* boundary-level annotations and *implicit* image-level object category annotations. From the former, the system will obtain direct information about image cues indicative of generic foreground object boundaries. From the latter, it will learn object-like features across a wide spectrum of object types—but *without* being told where those objects’ boundaries are.

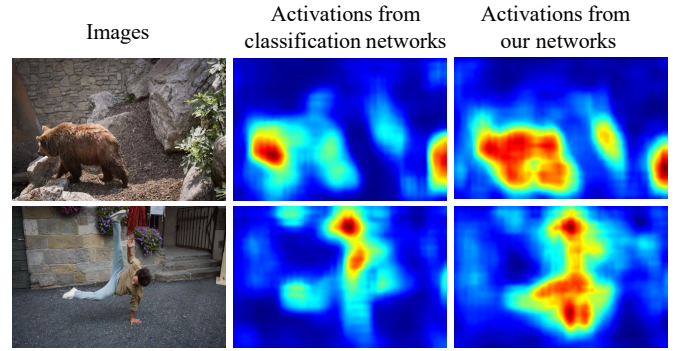


Fig. 2: Activation maps from a network (VGG [40]) trained for the classification task and our network which is fine-tuned with explicit dense foreground labels. We see that the classification network has already learned image representations that have some notion of objectness, but with poor “over”-localization. Our network deepens the notion of objectness to pixels and captures fine-grained cues about boundaries (best viewed on pdf).

To this end, we propose to train a fully convolutional deep neural network for the foreground-background object labeling task. We initialize the network using a powerful generic image representation learned from millions of images labeled by their object category, but lacking any foreground annotations. Then, we fine-tune the network to produce dense binary segmentation maps, using relatively few images with pixel-level annotations originating from a small number of object categories.

Since the pretrained network is trained to recognize thousands of objects, we hypothesize that its image representation has a strong notion of objectness built inside it, even though it never observes *any* segmentation annotations. Meanwhile, by subsequently training with explicit dense foreground labels, we can steer the method to fine-grained cues about boundaries that the standard object classification networks have no need to capture. This way, even if our model is trained with a limited number of object categories having pixel-level annotations, we expect it to learn generic representations helpful to pixel objectness.

Specifically, we adopt a deep network structure [4] originally designed for multi-class semantic segmentation. We initialize it with weights pre-trained on ImageNet, which provides a representation equipped to perform image-level classification for some 1,000 object categories. Next, we take a modestly sized semantic segmentation dataset, and transform its dense semantic masks into binary object vs. background masks, by fusing together all its 20 categories into a single supercategory (“generic object”). We then train the deep network (initialized for ImageNet object classification) to perform well on the dense foreground pixel labeling task. Our model supports end-to-end training.

To illustrate this synergy, Fig. 2 shows activation maps from a network trained for ImageNet classification (middle) and from our network (right), by summing up feature responses from each filter in the last convolutional layer (pool5) for each spatial location. Although networks trained on a classification task never observe any segmentations, they can show high activation responses when object parts are present and low activation responses to stuff-like regions such as rocks and roads. Since the classification networks are trained with thousands of object categories, their activation responses are rather general. However, they are responsive to only fragments of the

objects. After training with explicit dense foreground labels, our network is able to extend high activation responses from discriminative object parts to the entire object.

For example, in Fig. 2, the classification network only has a high activation response on the bear’s head, whereas our pixel objectness network has a high response on the entire bear body; similarly for the person. This supports our hypothesis that networks trained for classification tasks contain a reasonable but incomplete basis for objectness, despite lacking any spatial annotations. By subsequently training with explicit dense foreground labels, we can steer towards fine-grained cues about boundaries that the standard object classification networks have no need to capture.

Model architecture: We adapt the widely used image classification model VGG-16 network [40] into a fully convolutional network by transforming its fully connected layers into convolutional layers [3], [4]. This enables the network to accept input images of any size and also produce corresponding dense output maps. The network comprises of stacks of convolution layers with max-pooling layers in between. All convolution filters are of size 3×3 except the last convolution layer which comprises 1×1 convolutions. Each convolution layer is also followed by a “relu” non-linearity before being fed into the next layer. We remove the 1000-way classification layer from VGG-net and replace it with a 2-way layer that produces a binary mask as output. The loss is the sum of cross-entropy terms over each pixel in the output layer. See appendix for more details.

The VGG-16 network consists of five max pooling layers. While well suited for classification, this leads to a $32 \times$ reduction in the output resolution compared to the original image. In order to achieve more fine-grained pixel objectness map, we apply the “hole” algorithm proposed in [4]. In particular, this algorithm modifies the VGG-16 architecture by replacing the subsampling in the last two max-pooling layers with atrous convolution [4]. This method is parameter free and results in only a $8 \times$ reduction in the output resolution. We then use bilinear interpolation to recover a foreground map at the original resolution.

Training details: To generate the explicit boundary-level training data, we rely on the 1,464 PASCAL 2012 segmentation training images [38] and the additional annotations of [41], for 10,582 total training images. The 20 object labels are discarded, and mapped instead to the single generic “object-like” (foreground) label for training. We train our model using the Caffe implementation of [4]. We optimize with stochastic gradient with a mini-batch size of 10 images. A simple data augmentation through mirroring the input images is also employed. A base learning rate of 0.001 with a $1/10$ th slow-down every 2000 iterations is used. We train the network for a total of 10,000 iterations; total training time was about 8 hours on a modern GPU.

B. Leveraging Pixel Objectness

Dense pixel objectness has many applications. We are especially interested in how it can assist in image retrieval and content-aware image retargeting. We next present strategies for each task that leverage pixel objectness.

Object-aware image retrieval: First, we consider how pixel objectness foregrounds can assist in image retrieval. A retrieval system accepts a query image containing an object, and then the system returns a ranked list of images that contain the same object. This is a valuable application, for example, to allow object-based online product finding. Typically retrieval systems extract image features from the entire query image. This can be problematic, however, because it might retrieve images with similar background, especially when the object of interest is relatively small. We aim to use pixel objectness to restrict the system’s attention to the foreground object(s) as opposed to the entire image.

To implement the idea, we first run pixel objectness. In order to reduce false positive segmentations, we keep the largest connected foreground region if it is larger than 6% of the overall image area. Then we crop the smallest bounding box enclosing the foreground segmentation and extract features from the entire bounding box. If no foreground is found (which occurs in roughly 17% of all images), we extract image features from the entire image. The method is applied to both the query and database images. To rank database images, we explore two image representations. The first one uses only the image features extracted from the bounding box, and the second concatenates the features from the original image with those from the bounding box.

Foreground-aware image retargeting: As a second application, we explore how pixel objectness can enhance image retargeting. The goal is to adjust the aspect ratio or size of an image without distorting its important visual concepts. We build on the popular Seam Carving algorithm [42], which eliminates the optimal irregularly shaped path, called a seam, from the image via dynamic programming. In [42], the energy is defined in terms of the image gradient magnitude. However, the gradient is not always a sufficient energy function, especially when important visual content is non-textured or the background is textured.

Our idea is to protect semantically important visual content based on foreground segmentation. To this end, we consider a simple adaption of the Seam Carving. We define an energy function based on high-level semantics rather than low-level image features alone. Specifically, we first predict pixel objectness, and then we scale the gradient energy g within the foreground segment(s) by $(g + 1) \times 2$.

IV. RESULTS

We evaluate pixel objectness compared to 12 recent methods in the literature, and examine its utility for the two applications presented above.

Datasets: We use three challenging datasets:

- **MIT Object Discovery:** This challenging dataset consists of Airplanes, Cars, and Horses [20]. It is most commonly used to evaluate weakly supervised segmentation methods. The images were primarily collected using internet search and the dataset comes with per-pixel ground truth segmentation masks.
- **ImageNet-Localization:** We conduct a large-scale evaluation of our approach using ImageNet [43] (~ 1 M images

with bounding boxes, 3,624 classes). The diversity of this dataset lets us test the generalization abilities of our method.

- **ImageNet-Segmentation:** This dataset contains 4,276 images from 445 ImageNet classes with pixel-wise ground truth from [44].

Baselines: We compare to the following state-of-the-art methods:

- **Saliency Detection:** We compare to two salient object detection methods [8], [11], selected for their efficiency and state-of-the-art performance. Both methods are designed to produce a complete segmentation of the prominent object (vs. localized fixation maps) and output continuous saliency maps, which are then thresholded by per image mean to obtain the segmentation.³
- **Object Proposals:** We also compare with the state-of-the-art multiscale combinatorial grouping (MCG) algorithm [14] which outputs a ranked list of generic object segmentation proposals. The top ranked proposal in each image is taken as the final foreground segmentation for evaluation.
- **Weakly supervised joint-segmentation methods:** These approaches rely on weak supervision which comes in the form of prior knowledge that all images in a given collection share a common object category [20], [29], [45], [27], [28], [32], [37]. Note that our method lacks this additional supervision.

Evaluation metrics: Depending on the dataset, we use: (1)

Jaccard Score: Standard intersection-over-union (IoU) metric between predicted and ground truth segmentation masks and (2) **BBox-CorLoc Score:** Percentage of objects correctly localized with a bounding box according to PASCAL criterion (i.e. $\text{IoU} > 0.5$) used in [32], [26].

For MIT and ImageNet-Segmentation, we use the segmentation masks and evaluate using the Jaccard score. For ImageNet-Localization, since no large scale segmentation ground truth exists for imageNet, we follow the setup described in [32], [37], and consider all images with bounding box annotations available and evaluate using the BBox-CorLoc metric. We evaluate our segmentation masks against the bounding boxes annotations, using a tight bounding box around the foreground segmentation predicted by our method.

A. Foreground object segmentation results

MIT Object Discovery: First we present results on the MIT dataset [20]. We do separate evaluation on the complete dataset and also a subset defined in [20]. We compare our method with 9 existing state-of-the-art methods including saliency detection [8], [11], object proposal generation [14] and joint-segmentation [20], [29], [45], [27], [28], [37]. We compare with author-reported results for the joint-segmentation baselines, and use software provided by the authors for the saliency and object proposal baselines.

Table I shows the results. Our proposed method significantly advances state-of-the-art. It outperforms all existing methods

Methods	MIT dataset (subset)			MIT dataset (full)		
	Airplane	Car	Horse	Airplane	Car	Horse
# Images	82	89	93	470	1208	810
Joint Segmentation						
Joulin et al. [45]	15.36	37.15	30.16	n/a	n/a	n/a
Joulin et al. [27]	11.72	35.15	29.53	n/a	n/a	n/a
Kim et al. [28]	7.9	0.04	6.43	n/a	n/a	n/a
Rubinstein et al. [20]	55.81	64.42	51.65	55.62	63.35	53.88
Chen et al. [29]	54.62	69.2	44.46	60.87	62.74	60.23
Jain et al. [37]	58.65	66.47	53.57	62.27	65.3	55.41
Saliency						
Jiang et al. [11]	37.22	55.22	47.02	41.52	54.34	49.67
Zhang et al. [8]	51.84	46.61	39.52	54.09	47.38	44.12
Top Object Proposal						
MCG [14]	32.02	54.21	37.85	35.32	52.98	40.44
Ours	66.43	85.07	60.85	66.18	84.80	64.90
Rel. gain over best	13%	28%	14%	6%	30%	8%

TABLE I: Comparison with state-of-the-art methods on MIT Object Discovery dataset. Our method outperforms several state-of-the-art methods for saliency detection, object proposal generation, and joint segmentation. (Metric: Jaccard score). The last row shows the percentage improvement between our method and the next best method.

with absolute gains ranging from 4% to 19% over the best baseline, and relative gains up to 30%. The saliency detection and object proposal methods are substantially weaker than our method. This highlights the value of our end-to-end learning framework for effectively segmenting foreground objects.

Our gains over the joint segmentation methods are arguably even more impressive because our proposed model simply segments a single image at a time—no weak supervision!—and still substantially outperforms all weakly supervised joint segmentation techniques. We stress that in addition to the weak supervision in form of segmenting common object, the previous best performing method [37] also makes use of a pre-trained deep network; hence we use strictly less total supervision than [37] yet still perform better.

Furthermore, most joint segmentation methods involve expensive steps such as dense correspondences [20] or region matching [37] which can take up to hours even for a modest collection of 100 images. In contrast, our method directly outputs the final segmentation in a single forward pass over the deep network and takes only 0.6 seconds per image for complete processing.

ImageNet-Localization: Next we present our segmentation results on ImageNet-Localization dataset. This involves testing our method on about 1 million images from 3,624 object categories. This also lets us test how generalizable our method is to unseen categories, i.e., those for which the method sees no foreground examples during training.

Table II shows the results. When doing the evaluation over all categories, we compare our method with results reported by 5 existing techniques. We see that our method significantly improves the state-of-the-art. The saliency and object proposal methods result in much poorer segmentations. Our method also significantly outperforms the joint segmentation approaches [32], [37], which are the current best performing methods on this dataset. In terms of the actual number of images, our gains translate into correctly segmenting 42,900 more images than [37] (which, like us, leverages ImageNet features) and 83,800 more images than [32]. This reflects the overall magnitude of our gains over state-of-the-art methods.

Does our learned segmentation model only recognize fore-

³This thresholding strategy was chosen because it gave the best results.

ImageNet-Localization dataset		
All	# Classes	# Images
	3,624	939,516
Non-PASCAL	# Classes	# Images
	3,149	810,219

Methods	BBox-CorLoc	
	All	Non-Pascal
Top-Objectness (Alexe) [46]	37.42	n/a
Tang et al. [32]	53.20	n/a
Jain et al. [37]	57.64	n/a
Saliency [11]	41.28	39.35
Top-Objectness (MCG) [14]	42.23	41.15
Ours	62.12	60.18

TABLE II: Comparison with state-of-the-art methods on ImageNet-Localization dataset. Our proposed segmentation model outperforms several state-of-the-art methods and also generalizes very well to unseen object categories. (Metric: BBox-CorLoc).

ground objects that it has seen during training, or can it generalize to unseen object categories? Intuitively, ImageNet has such a large number of diverse categories that this gain in performance would not have been possible if our method was only over-fitting to the 20 seen PASCAL object categories. To empirically verify this intuition, we next exclude those ImageNet categories which are directly related to the PASCAL objects, by matching the two datasets’ synsets. This results in a total of 3,149 categories which are exclusive to ImageNet (“Non-PASCAL”). See Table II (top) for the data statistics.

We see only a very marginal drop in performance; our method still significantly outperforms both the saliency and object proposal baselines. This is an important result, because during training the segmentation model *never saw any dense object masks for images in these categories*. Bootstrapping from the pretrained weights of the VGG-classification network, our model is able to learn a transformation between its prior belief on what looks like an object to complete dense foreground segmentations.

ImageNet-Segmentation: Finally, we measure the pixel-wise segmentation quality on a large scale. For this we use the ground truth masks provided by [44] for 4,276 images from 445 ImageNet categories. For this dataset the current best results are due to the segmentation propagation approach of [44]. Our method again improves the state-of-the-art significantly, with a Jaccard score of 64.22 as opposed to 57.3 reported by [44]. This shows that our method not only generalizes to thousands of object categories but also produces high quality object segmentations.

Qualitative results: Fig. 3 shows qualitative results for the ImageNet dataset for both PASCAL and Non-PASCAL categories. Pixel objectness accurately segments foreground objects from both sets. The examples from the Non-PASCAL categories highlight its strong generalization capabilities. We are able to segment objects across all scales and appearance variations, including multiple objects within an image. The bottom few examples show our method’s remarkable ability to segment even man-made objects, which are especially distinct from the kind of objects in PASCAL (see appendix for more examples). The bottom row shows some failure cases. We observe that our model has more difficulty in segmenting scene-centric images. It is understandable because in most scene-centric images, the entire scene is of primary importance

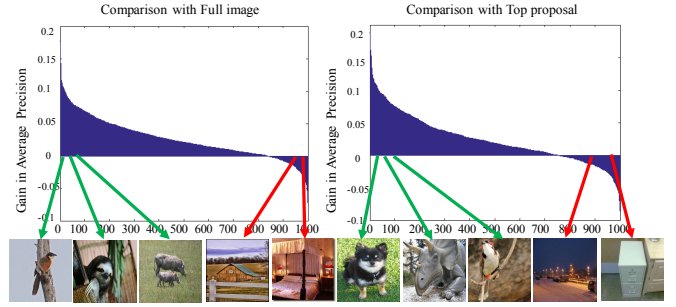


Fig. 4: We show the *gain* in average precision per object class between our method and the baselines (Full image on the left, and Top proposal on the right). Green arrows indicate example object classes for which our method performs better and red arrows indicate object classes for which the baselines perform better. Note our method excels at retrieving natural objects but can fail for scene-centric classes.

and it is more difficult to clearly identify foreground objects.

B. Impact on downstream applications

Next we report results leveraging pixel objectness for two downstream tasks.

Object-aware image retrieval

First we consider the object-based image retrieval task defined in Sec. III-B. We use the ILSVRC2012 [47] validation set, which contains 50K images and 1,000 object classes, with 50 images per class. As an evaluation metric, we use mean average precision (mAP). We extract VGGNet [40] features and use cosine distance to rank retrieved images.

We compare with two baselines 1) **Full image**, which ranks images based on features extracted from the entire image, and 2) **Top proposal** (TP), which ranks images based on features extracted from the top ranked MCG [14] proposal. For our method and the Top proposal baseline, we examine two image representations. The first directly uses the features extracted from the region containing the foreground or the top proposal (denoted **FG**). The second representation concatenates the extracted features with the image features extracted from the entire image (denoted **FF**).

Table III shows the results. Our method with FF yields the best results. Our method outperforms both baselines for many ImageNet classes. Figure 4 looks more closely at the distribution of our method’s gains in average precision per class. We observe that our method performs extremely well on object-centric classes such as animals, but has limited improvement upon the baseline on classes that are scene-centric (lakeshore, seashore etc.) since separating background from a scene-centric image has limited effect. To verify our hypothesis, we isolate the results on the first 400 object classes of ImageNet, which contain mostly object-centric classes, as opposed to scene-centric objects. On those first 400 object classes, our method outperforms both baselines by a larger margin. This demonstrates the value of our method at retrieving objects, which often contain diverse background and so naturally benefit more from accurate pixel objectness.

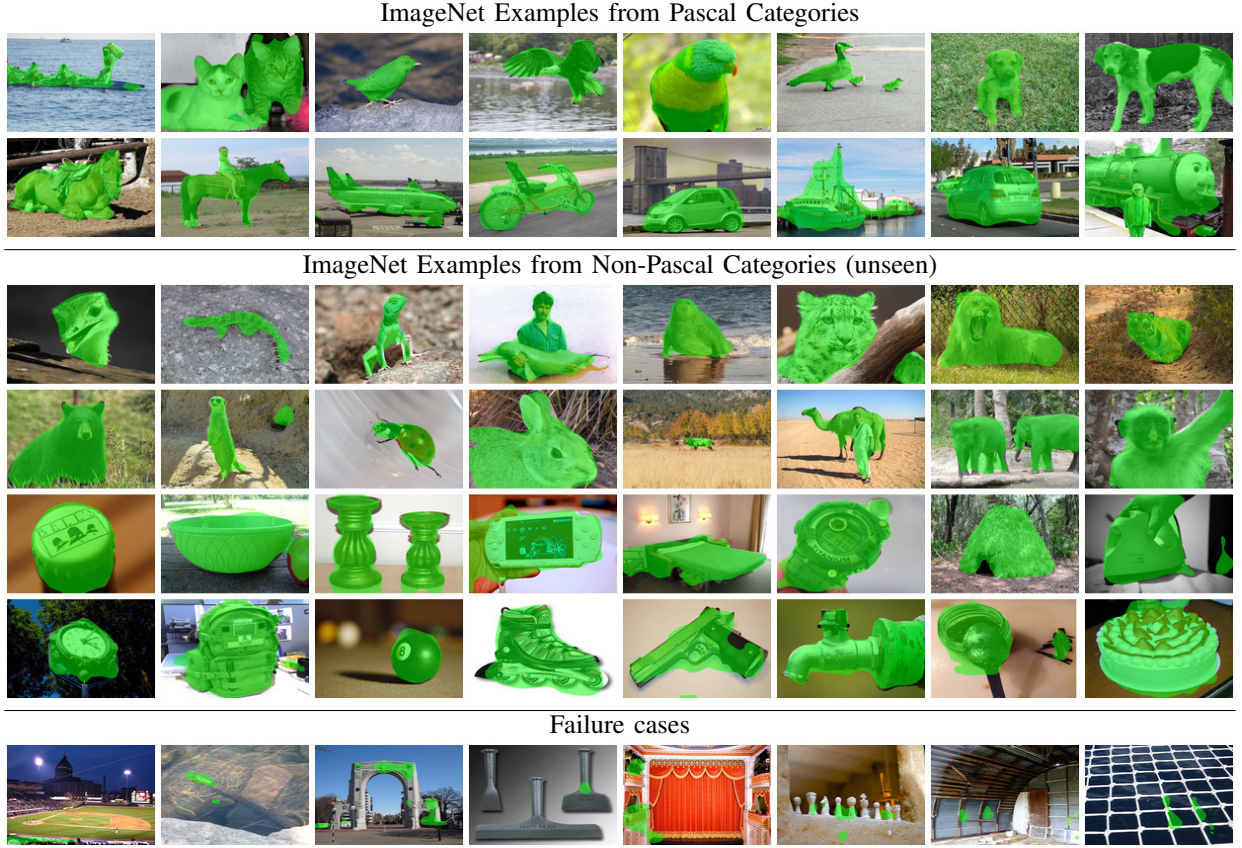


Fig. 3: Qualitative results: We show qualitative results on images belonging to PASCAL (top) and Non-PASCAL (middle) categories. Our segmentation model generalizes remarkably well even to those categories which were unseen in any foreground mask during training (middle rows). Typical failure cases (bottom) involve scene-centric images where it is not easy to clearly identify foreground objects (best viewed on pdf).

Method	Ours(FG)	Ours(FG)	Full Img	TP (FF) [14]	TP (FG) [14]
All	0.3342	0.3173	0.3082	0.3102	0.2092
Obj-centric	0.4166	0.4106	0.3695	0.3734	0.2679

TABLE III: Object-based image retrieval performance on ImageNet. We report average precision on the entire validation set, and on the first 400 categories, which are mostly object-centric classes.

To further understand the superior performance of our proposed method, we show the Top-5 nearest neighbors for both our method and the Full image baseline in Figure 5. In the first example (first and second rows), the query image contains a small bird standing on a bird house. Our method is able to segment the bird and retrieves relevant images that also contain birds. The baseline, on the contrary, has noisier retrievals due to mixing the background. The last two rows show a case where, at least according to ImageNet labels, our method fails. Our method segments the person as foreground, and then retrieves images containing a person from different scenes, whereas the baseline focuses on the entire image and retrieves similar scenes.

Foreground-aware image retargeting

Next, we show how to enhance Seam Carving retargeting with pixel objectness predictions. We use a random subset of 500 images from the 2014 Microsoft COCO Captioning Challenge Testing Images [48] for experiments.

Figure 6 shows example results. For reference, we also compare with the original Seam Carving retargeting algorithm [42] that uses image gradients as the energy function. Both



Fig. 5: We show the top 5 nearest neighbors for both the Full Img baseline and our method. Baseline row shows the original query (left) with our predicted fg in green. Our row shows the images cropped to our fg estimates (query and neighbors). Our method focuses on the foreground to retrieve images that contain the same object. Second example is a failure case on a scene-centric image; our method segments a person as foreground and retrieves images that also contain persons (best viewed on pdf).

methods are instructed to resize the source image to 2/3 of its original size, in terms of height and width.

Our method preserves important objects in the images thanks to the proposed foreground segmentation. The baseline produces images with important objects distorted, because gradient strength is an inadequate indicator for perceived



Fig. 6: We show original images with predicted foreground in green (prediction, top row), retargeted images produced by our method (Ours, middle row) and retargeted images produced by the Seam Carving based on gradient energy [42] (SC, bottom row). Our method successfully preserves the important visual content (e.g., train, bus, human and dog) while reducing the content of the background. The rightmost column is a failure case. See appendix for more examples (best viewed on pdf).

content, especially when background is textured.

To quantify the results over all 500 images, we perform a human study on Amazon Mechanical Turk. We present image pairs produced by our method and the baseline in arbitrary order and ask workers to rank which image is more likely to have been manipulated by a computer. Each image pair is evaluated by three different workers. Workers found that 38.53% of the time images produced by our method are more likely to have been manipulated by a computer, 48.87% for the baseline; both methods tie 12.60% of the time. Thus, human evaluation with non-experts demonstrates that our method outperforms the baseline. In addition, we also ask a vision expert familiar with the task of image retargeting—but not involved in this project—to score the 500 image pairs with the same interface as the crowd workers. Out of 500 images, the vision expert found our method performs better for 78% of the images, baseline is better for 13% of the images, and both methods tie for 9% images. This further confirms that our foreground prediction can enhance image retargeting by defining a more semantically meaningful energy function.

V. CONCLUSIONS

We proposed an end-to-end learning framework for segmenting generic foreground objects in images. Our results demonstrate its effectiveness, with significant improvements over the state-of-the-art on multiple datasets. Our results also show that pixel objectness generalizes very well to thousands of unseen object categories. The foreground segmentations produced by our model also proved to be highly effective in improving the performance of image-retrieval and image-retargeting tasks, which helps illustrate the real-world demand for high-quality, single image, non-interactive foreground segmentations.

Acknowledgements: This research is supported in part by ONR YIP N00014-12-1-0754.

VI. APPENDIX

A. CNN Architecture (Sec. III-A in the main text)

Here we provide more details of the fully convolutional architecture that was employed to train our model for pixel objectness.

Notations:

- 1) **Convolution layers:** conv x - y denotes a convolution layer with $x \times x$ kernels and y channels, a stride of 1 was used everywhere.
- 2) **Max Pooling:** maxpool denotes a max-pooling layer with KS as kernel size and stride S.
- 3) **Non Linearity:** A relu non-linear activation function was used after each convolution layer.
- 4) **Dropout:** Dropout regularization was used in the last layers with a ratio of 0.5.

Architecture:

- Input Image: 3-channel RGB ($3 \times 321 \times 321$)
- conv3-64 \rightarrow relu \rightarrow conv3-64 \rightarrow relu \rightarrow maxpool (KS:3, S:2)
- conv3-128 \rightarrow relu \rightarrow conv3-128 \rightarrow relu \rightarrow maxpool (KS:3, S:2)
- conv3-256 \rightarrow relu \rightarrow conv3-256 \rightarrow relu \rightarrow conv3-256 \rightarrow relu \rightarrow maxpool (KS:3, S:2)
- conv3-512 \rightarrow relu \rightarrow conv3-512 \rightarrow relu \rightarrow conv3-512 \rightarrow relu \rightarrow maxpool (KS:3, S:1)
- conv3-512 \rightarrow relu \rightarrow conv3-512 \rightarrow relu \rightarrow conv3-512 \rightarrow relu \rightarrow maxpool (KS:3, S:1)
- conv3-1024 \rightarrow relu \rightarrow dropout (0.5) \rightarrow conv1-1024 \rightarrow relu \rightarrow dropout (0.5) \rightarrow conv1-2

Even though this architecture largely follows the standard VGG-16 [40] architecture, there are minor changes similar to [4] which enables us to obtain a higher resolution output map. This modified network was initialized using the VGG-16 pre-trained weights provided by [4].

B. More Qualitative Results of Pixel Objectness (Sec. IV-A in the main text)

We also show additional qualitative results from ImageNet dataset for our proposed pixel objectness model. Figure 7, 8 show the qualitative results for ImageNet images which belong to PASCAL categories. Our method is able to accurately segment foreground objects including cases with multiple foreground objects as well as the ones where the foreground objects are not highly salient.

Figure 9, 10, 11 show qualitative results for those ImageNet images which belong to the non-PASCAL categories. Even though trained only on foregrounds from PASCAL categories, our method generalizes surprisingly well. As can be seen, it can accurately segment foreground objects from completely disjoint categories, examples of which were never seen during training. Figure 12 shows more failure cases.

C. Foreground-aware image retargeting examples (Sec. IV-B in the main text)

We also present more foreground-aware image retargeting example results in Figure 13. Please refer to Sec. IV-B in the main paper for algorithmic details. Our method is able to preserve important objects in the images thanks to the proposed foreground segmentation method. The baseline produces images with important objects distorted, because gradient strength is not a good indicator for perceived content, especially when background is textured. We also present a few failure cases in the rightmost column. In the first example, our method is unsuccessful at predicting the skateboard as the foreground and therefore results in an image with skateboard distorted. In the second example, our method is able to detect and preserve all the people in the image. However, the background distortion creates artifacts that make the resulting image unpleasant to look at compared to the baseline. In the last example, our method misclassified the pillow as foreground and results in an image with an amplified pillow.

D. Amazon Mechanical Turk interface (Sec. IV-B in the main text)

We also show the interface we used to collect human judgement for image retargeting on Amazon Mechanical Turk. The two images produced by our algorithm and the baseline method are shown in arbitrary order to the workers. We instruct the workers to pick an image that is more likely to have been manipulated by a computer. If both images look like they have been manipulated by a computer, then pick the one that is manipulated more. The workers have three options to choose from: 1) The first image is more likely to have been manipulated by a computer; 2) The second image is more likely to have been manipulated by a computer; 3) Both images look equally real. See Figure 14 for the interface. Also refer to Sec. IV-B in the main paper for more discussions on these user study results.

REFERENCES

- [1] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Computer Vision (ICCV), 2015 IEEE International Conference on*, 2015.
- [2] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *International Conference on Computer Vision (ICCV)*, 2015.
- [3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CVPR*, Nov. 2015.
- [4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015. [Online]. Available: <http://arxiv.org/abs/1412.7062>
- [5] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *CVPR*. IEEE Computer Society, 2015, pp. 362–370. [Online]. Available: <http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#LiuHZWL15>
- [6] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," *CoRR*, vol. abs/1510.02927, 2015. [Online]. Available: <http://arxiv.org/abs/1510.02927>
- [7] J. Pan, K. McGuinness, E. Sayrol, N. O'Connor, and X. Giro-i Nieto, "Shallow and deep convolutional networks for saliency prediction," 2016.
- [8] J. Zhang and S. Sclaroff, "Saliency detection: a boolean map approach," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2013.
- [9] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *CVPR*, 2011, pp. 409–416.
- [10] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012, pp. 733–740.
- [11] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *ICCV*, 2013.
- [12] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *PAMI*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
- [13] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *PAMI*, vol. 34, no. 7, pp. 1312–1328, 2012.
- [14] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *CVPR*, 2014.
- [15] P. Krähenbühl and V. Koltun, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V*. Cham: Springer International Publishing, 2014, ch. Geodesic Object Proposals, pp. 725–739. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10602-1_47
- [16] I. Endres and D. Hoiem, "Category independent object proposals," in *ECCV*, 2010.
- [17] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.
- [18] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013. [Online]. Available: <https://ivi.fnwi.uva.nl/isis/publications/2013/UijlingsIJCV2013>
- [19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.
- [20] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *CVPR*, 2013.
- [21] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut -interactive foreground extraction using iterated graph cuts," in *SIGGRAPH*, 2004.
- [22] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive Co-segmentation with Intelligent Scribble Guidance," in *CVPR*, 2010.
- [23] S. Jain and K. Grauman, "Predicting sufficient annotation strength for interactive foreground segmentation," in *ICCV*, 2013.
- [24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [25] B. Alexe, T. Deselaers, and V. Ferrari, "Classcut for unsupervised class segmentation," in *ECCV*, 2010.
- [26] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *International Journal of Computer Vision*, vol. 100, no. 3, pp. 275–293, September 2012.
- [27] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *CVPR*, 2012.
- [28] G. Kim, E. Xing, L. Fei Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *ICCV*, 2011.

- [29] X. Chen, A. Shrivastava, and A. Gupta, “Enriching Visual Knowledge Bases via Object Discovery and Segmentation,” in *CVPR*, 2014.
- [30] S. Vicente, C. Rother, and V. Kolmogorov, “Object cosegmentation,” in *CVPR*, 2011.
- [31] J. Serrat, A. Lopez, N. Paragios, and J. C. Rubio, “Unsupervised co-segmentation through region matching,” in *CVPR*, 2012.
- [32] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, “Co-localization in real-world images,” in *CVPR*, 2014.
- [33] E. Ahmed, S. Cohen, and B. Price, “Semantic object selection,” in *CVPR*, June 2014.
- [34] J. Yang, B. Price, S. Cohen, Z. Lin, and M.-H. Yang, “Patchcut: Data-driven object segmentation via local shape transfer,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [35] D. Kuettel and V. Ferrari, “Figure-ground segmentation by transferring window masks,” in *CVPR*, 2012.
- [36] M. Guillaumin, D. Küttel, and V. Ferrari, “ImageNet auto-annotation with segmentation propagation,” *International Journal of Computer Vision*, vol. 110, no. 3, pp. 328–348, 2014.
- [37] S. Jain and K. Grauman, “Active image segmentation propagation,” in *CVPR*, 2016.
- [38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010. [Online]. Available: <http://dx.doi.org/10.1007/s11263-009-0275-4>
- [39] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *ECCV*, 2014.
- [40] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [41] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, “Semantic contours from inverse detectors,” in *International Conference on Computer Vision (ICCV)*, 2011.
- [42] S. Avidan and A. Shamir, “Seam carving for content-aware image resizing,” in *ACM Transactions on graphics (TOG)*, vol. 26. ACM, 2007, p. 10.
- [43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [44] M. Guillaumin, D. Küttel, and V. Ferrari, “Imagenet auto-annotation with segmentation propagation,” *International Journal of Computer Vision*, vol. 110, no. 3, pp. 328–348, 2014.
- [45] A. Joulin, F. Bach, and J. Ponce, “Discriminative clustering for image co-segmentation,” in *CVPR*, 2010.
- [46] B. Alexe, T. Deselaers, and V. Ferrari, “What is an Object?” in *CVPR*, 2010.
- [47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 740–755.



Fig. 7: Qualitative results: We show example segmentations from ImageNet dataset obtained by our pixel objectness model. The segmentation results are shown with a green overlay. Our method is able to accurately segment foreground objects including cases where the objects are not highly salient. Best viewed in color.



Fig. 8: Qualitative results: We show example segmentations from ImageNet dataset obtained by our pixel objectness model on PASCAL Categories. The segmentation results are shown with a green overlay. Our method is able to accurately segment foreground objects including cases where the objects are not highly salient. Best viewed in color.

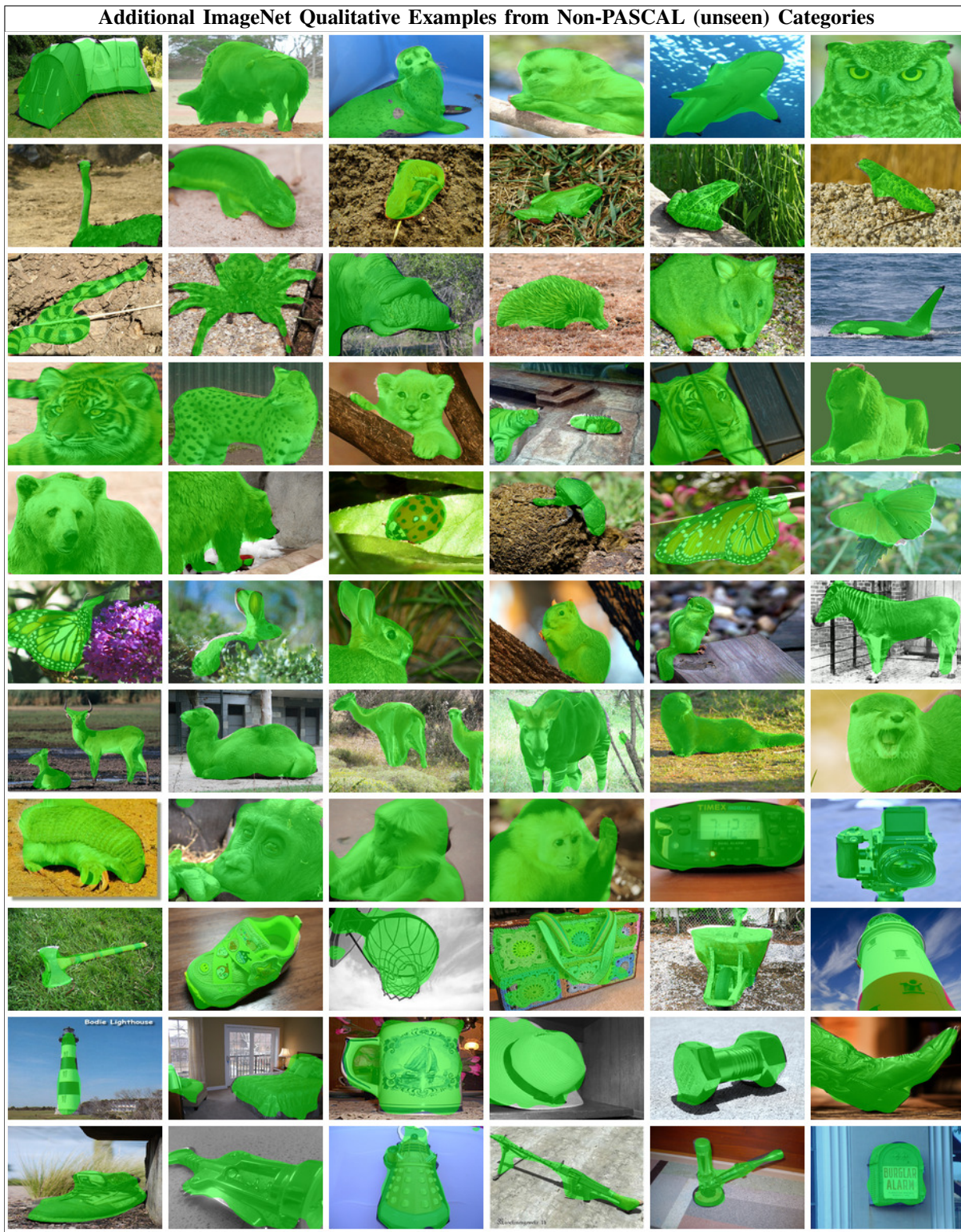


Fig. 9: Qualitative results: We show example segmentations from ImageNet dataset obtained by our pixel objectness model on Non-PASCAL Categories. The segmentation results are shown with a green overlay. Our method generalizes remarkably well and is able to accurately segment foreground objects even for those categories which were never seen during training. Best viewed in color.

Additional ImageNet Qualitative Examples from Non-PASCAL (unseen) Categories

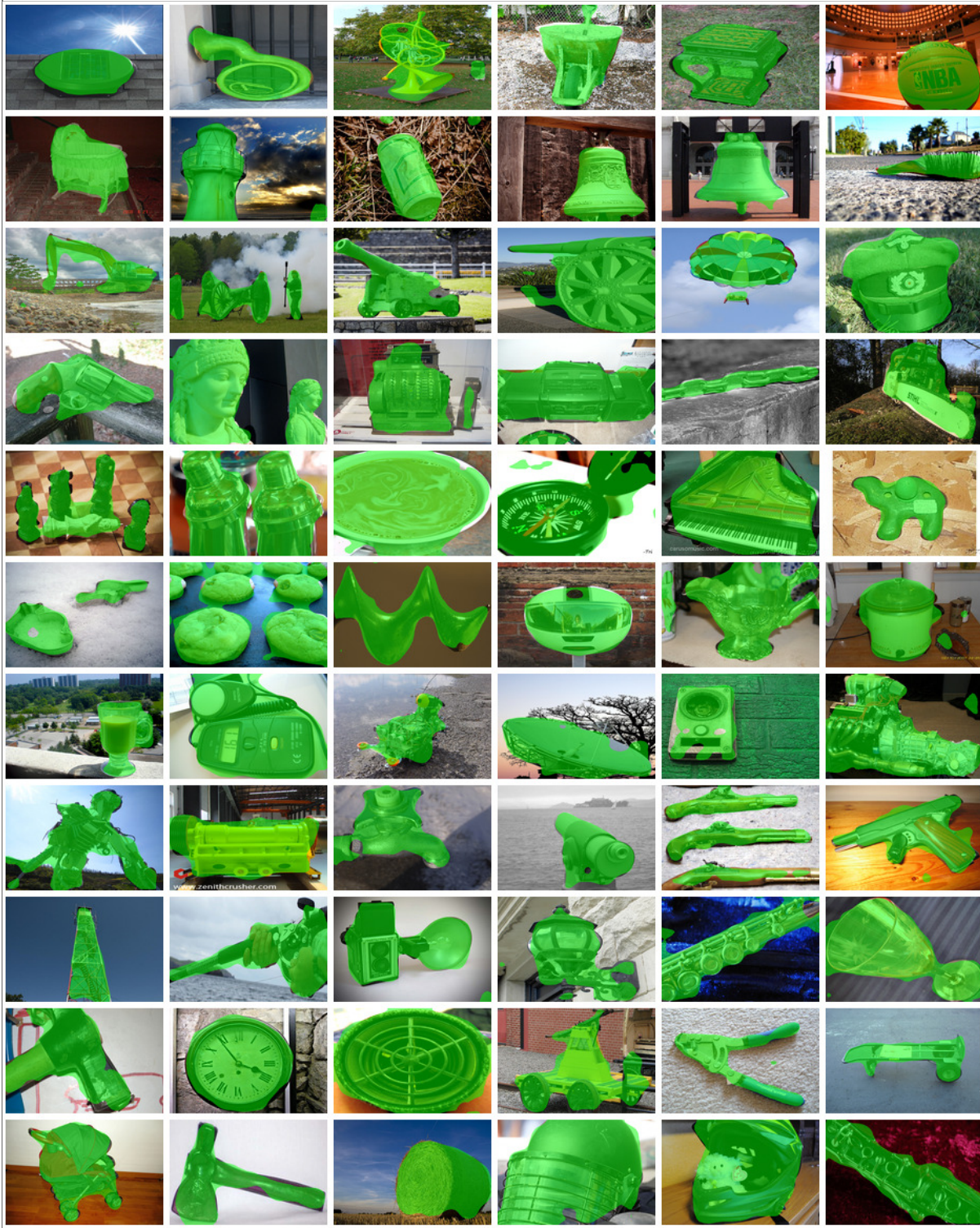


Fig. 10: Qualitative results: We show example segmentations from ImageNet dataset obtained by our pixel objectness model on Non-PASCAL Categories. The segmentation results are shown with a green overlay. Our method generalizes remarkably well and is able to accurately segment foreground objects even for those categories which were never seen during training. Best viewed in color.

Additional ImageNet Qualitative Examples from Non-PASCAL (unseen) Categories



Fig. 11: Qualitative results: We show example segmentations from ImageNet dataset obtained by our pixel objectness model on Non-PASCAL Categories. The segmentation results are shown with a green overlay. Our method generalizes remarkably well and is able to accurately segment foreground objects even for those categories which were never seen during training. Best viewed in color.

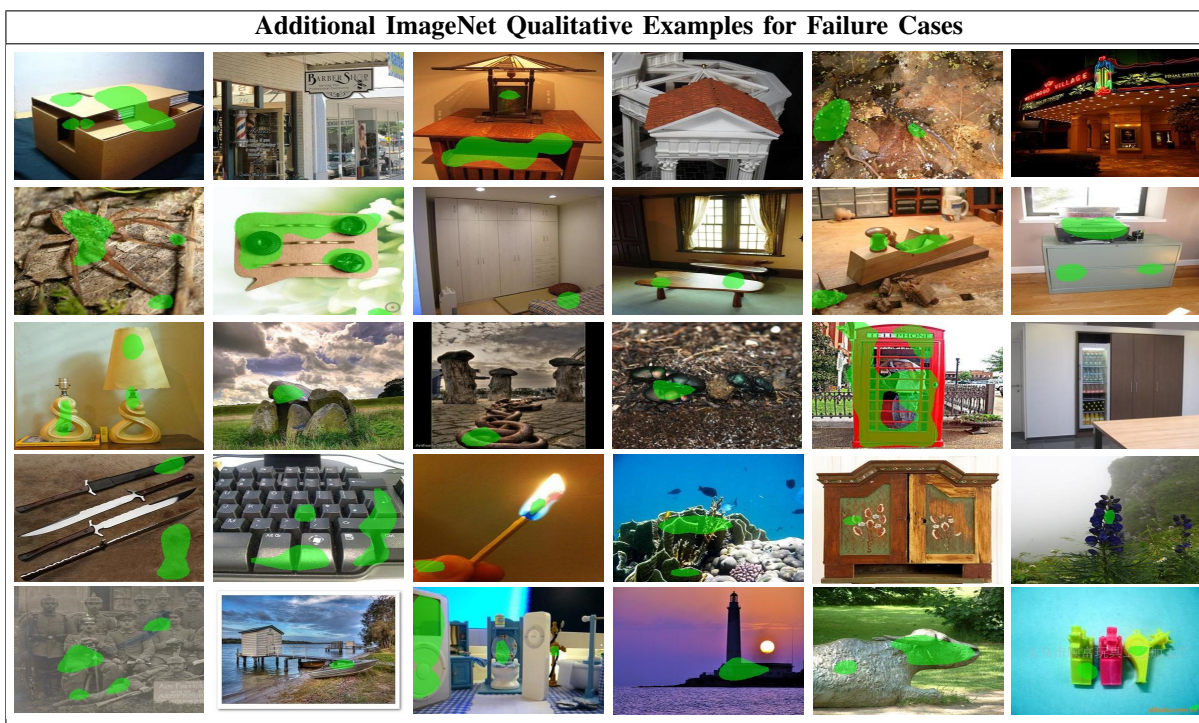


Fig. 12: Qualitative results: We show examples of failure cases from ImageNet dataset obtained by our pixel objectness model. The segmentation results are shown with a green overlay. Typical failure cases involve scene-centric images or images containing very thin objects. Best viewed in color.

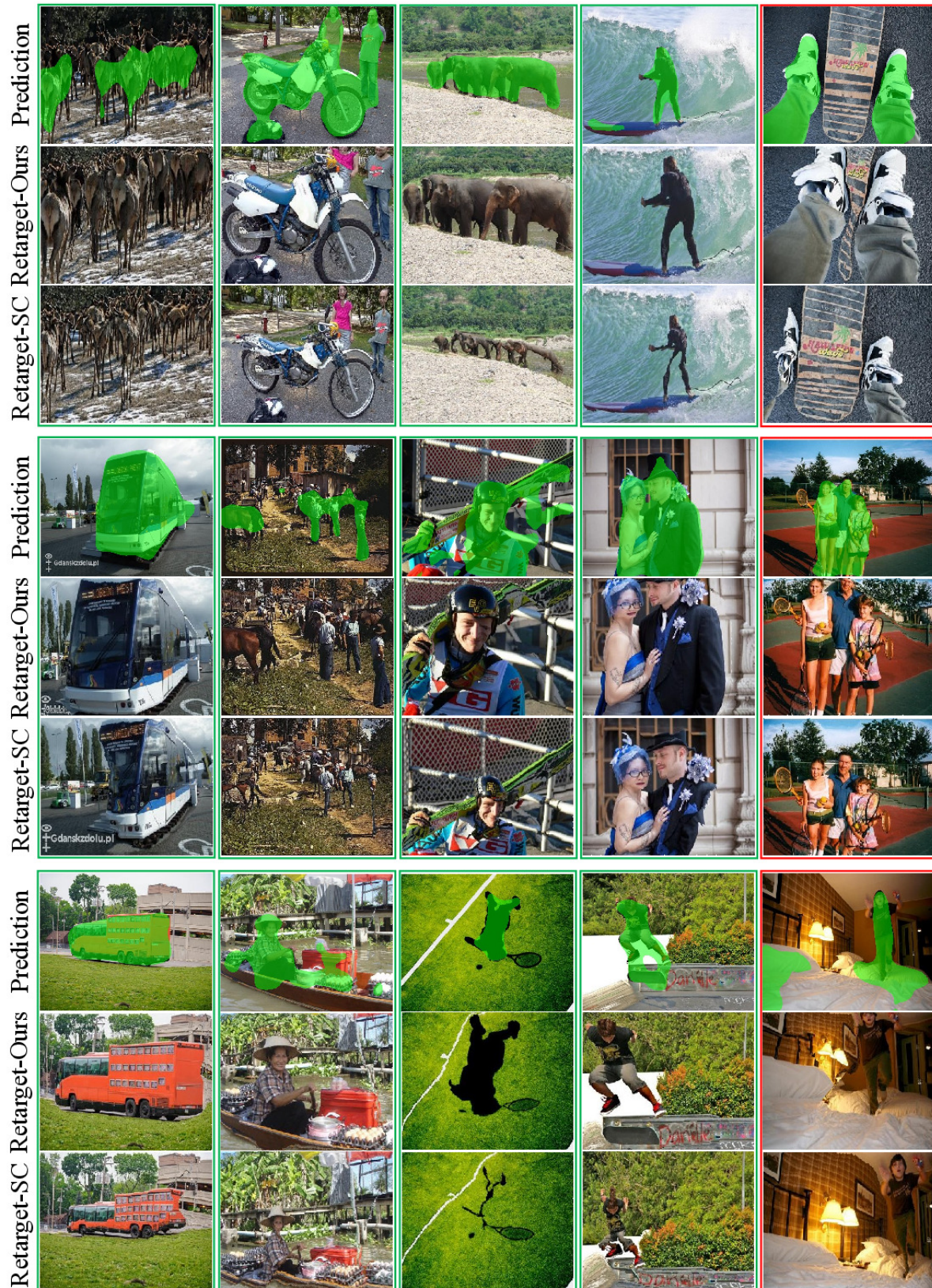



Fig. 13: We show more foreground-aware image retargeting example results. We show original images with predicted foreground in green (prediction, top row), retargeting images produced by our method (Retarget-Ours, middle row) and retargeting images produced by the Seam Carving based on gradient energy [42] (Retarget-SC, bottom row). Our method successfully preserves the important visual content while reducing the content of the background. We also present a few failure cases in the rightmost column. Best viewed in color.


Which image is more likely to have been manipulated by a computer?

Instructions
A computer has manipulated one of these images. Can you tell which image is more likely to have been manipulated by a computer? If both images look like they have been manipulated by a computer, then pick the image that you think is manipulated more.

Image1: which image is more likely to have been manipulated by a computer?



First Image



Second Image

Category:

- ☐ The first image is more likely to have been manipulated by a computer
- ☐ The second image is more likely to have been manipulated by a computer
- ☐ Both images look equally real

Fig. 14: Amazon Mechanical Turk interface used to collect human judgement for image retargeting. We ask workers to judge which image is more likely to have been manipulated by a computer. They have three options: 1) The first image is more likely to have been manipulated by a computer; 2) The second image is more likely to have been manipulated by a computer; 3) Both images look equally real.